Supplementary Information

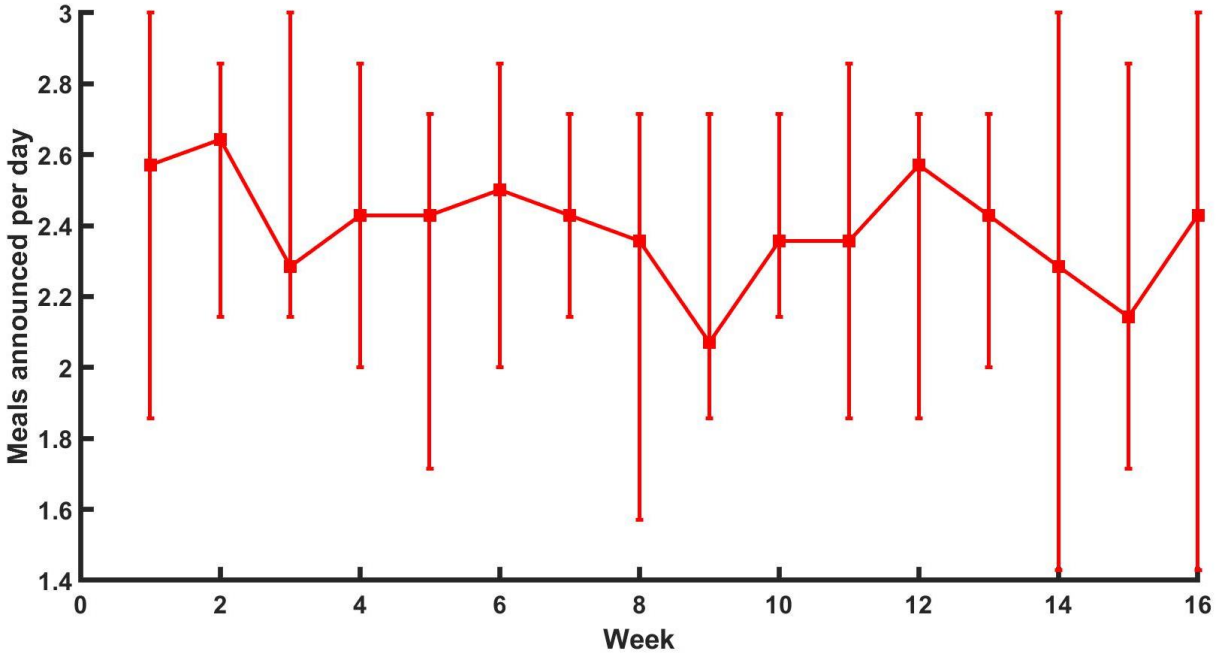


Figure S1. iBolusV2 app usage per day throughout the study duration (median and quartiles; n=14). Source data are provided as a Source Data file.

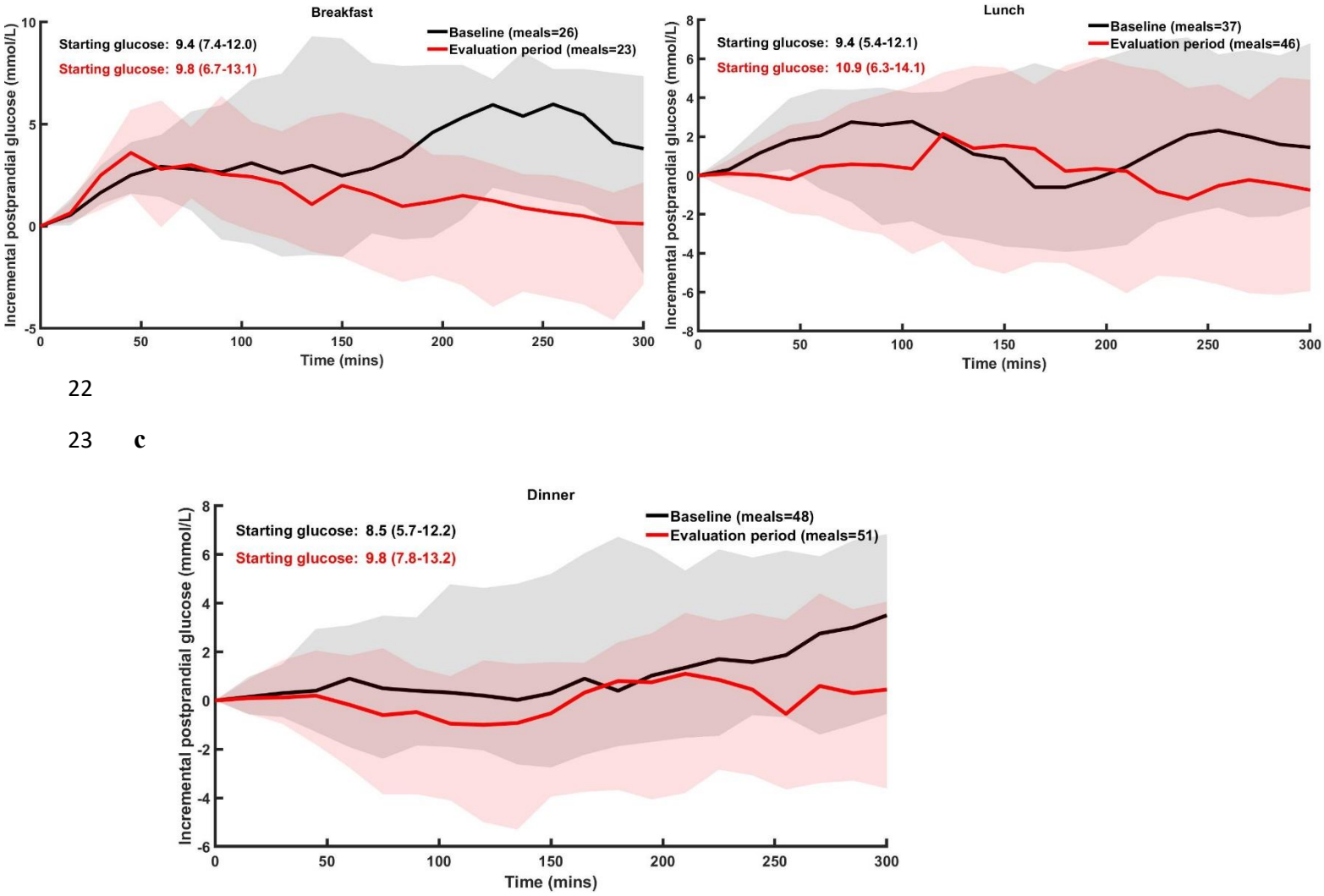


Figure S2. Postprandial incremental glucose levels (median and interquartile range) after high-fat meals (with no postprandial exercises) **(a)** breakfast meals, **(b)** lunch meals, and **(c)** dinner meals. Source data are provided as a Source Data file.

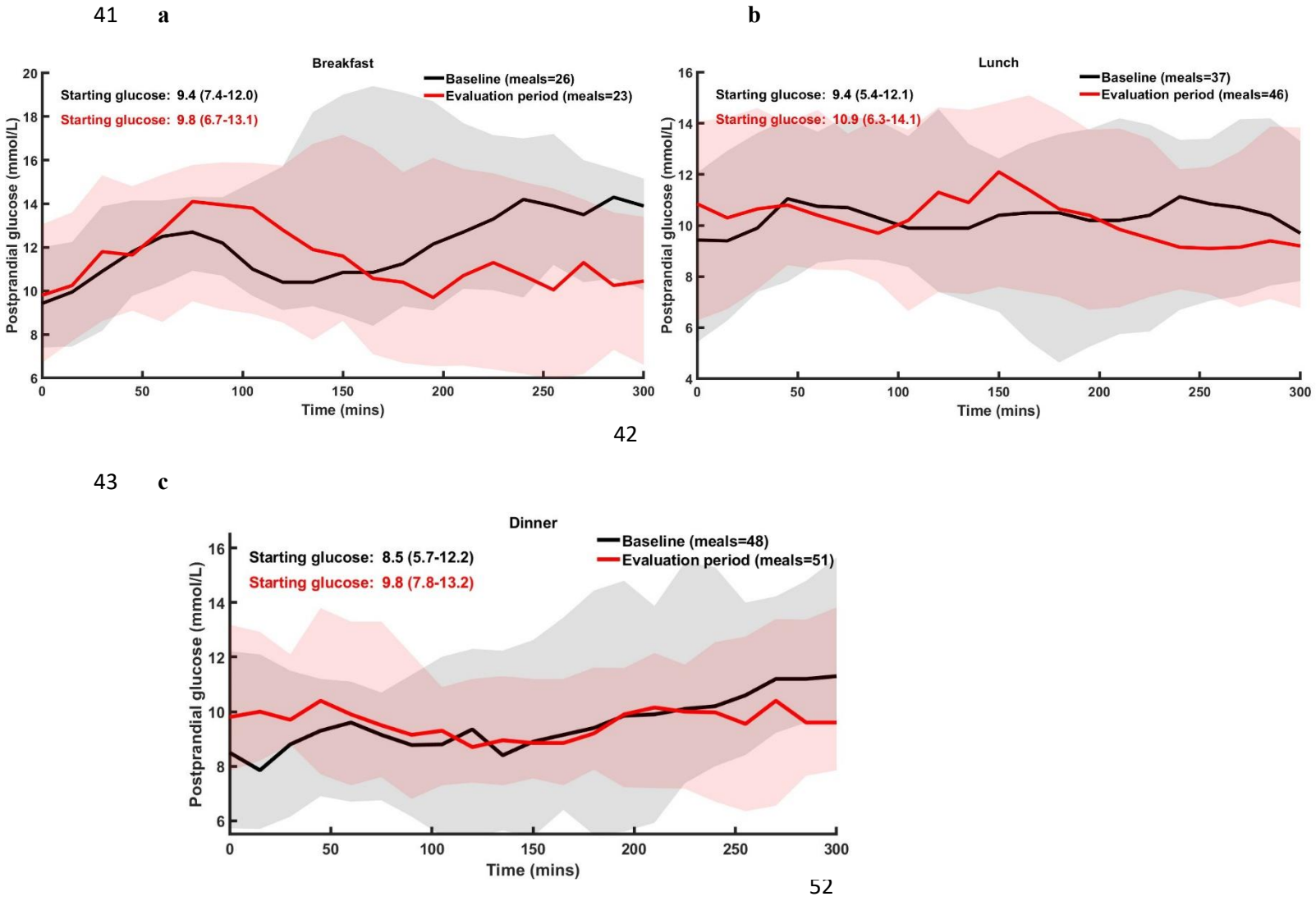


Figure S3. Postprandial glucose levels (median and interquartile range) after high-fat meals (with no postprandial exercises) **(a)** breakfast meals, **(b)** lunch meals, and **(c)** dinner meals. Source data are provided as a Source Data file.

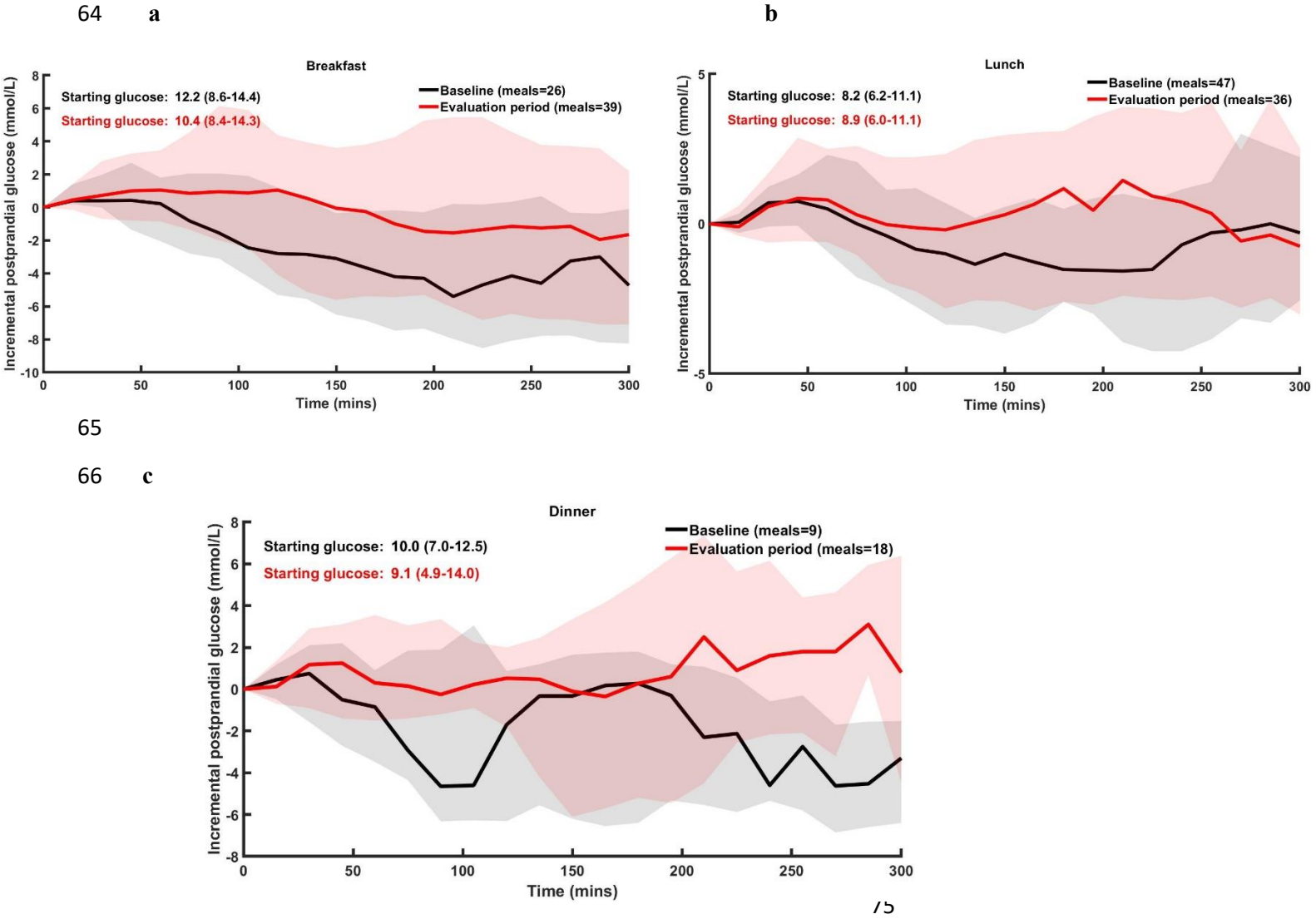


Figure S4. Postprandial incremental glucose levels (median and interquartile range) for meals followed by aerobic exercises **(a)** breakfast meals, **(b)** lunch meals, and **(c)** dinner meals. Source data are provided as a Source Data file.

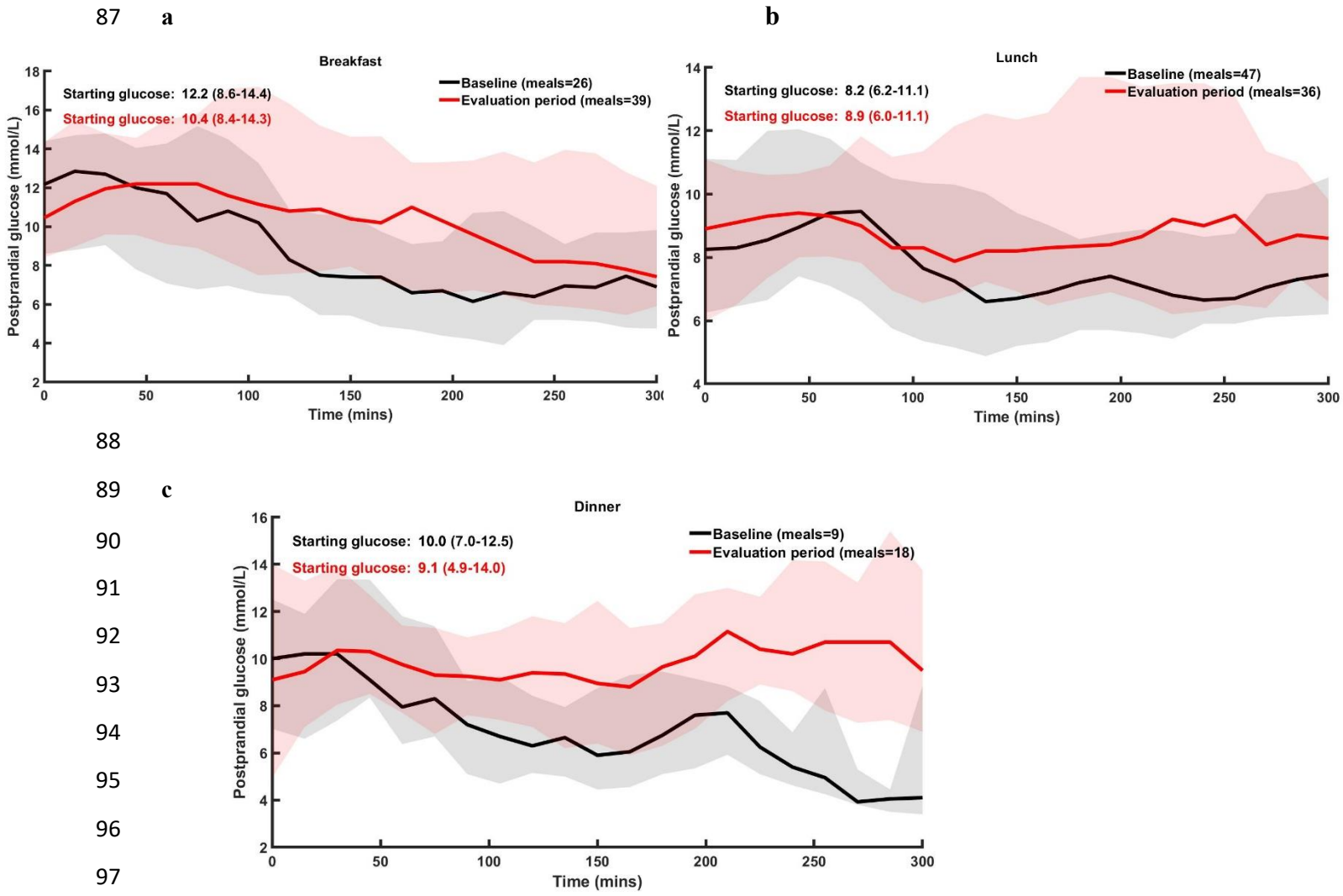
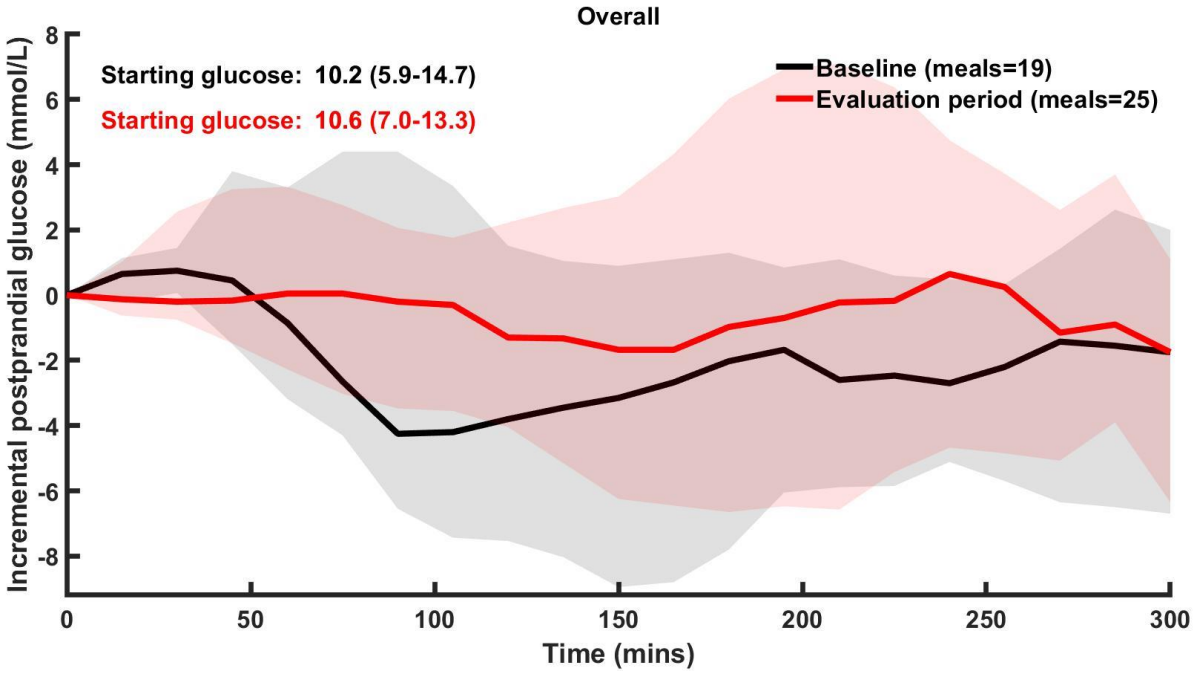


Figure S5. Postprandial glucose levels (median and interquartile range) for meals followed by aerobic exercises **(a)** breakfast meals, **(b)** lunch meals, and **(c)** dinner meals. Source data are provided as a Source Data file.

a



b

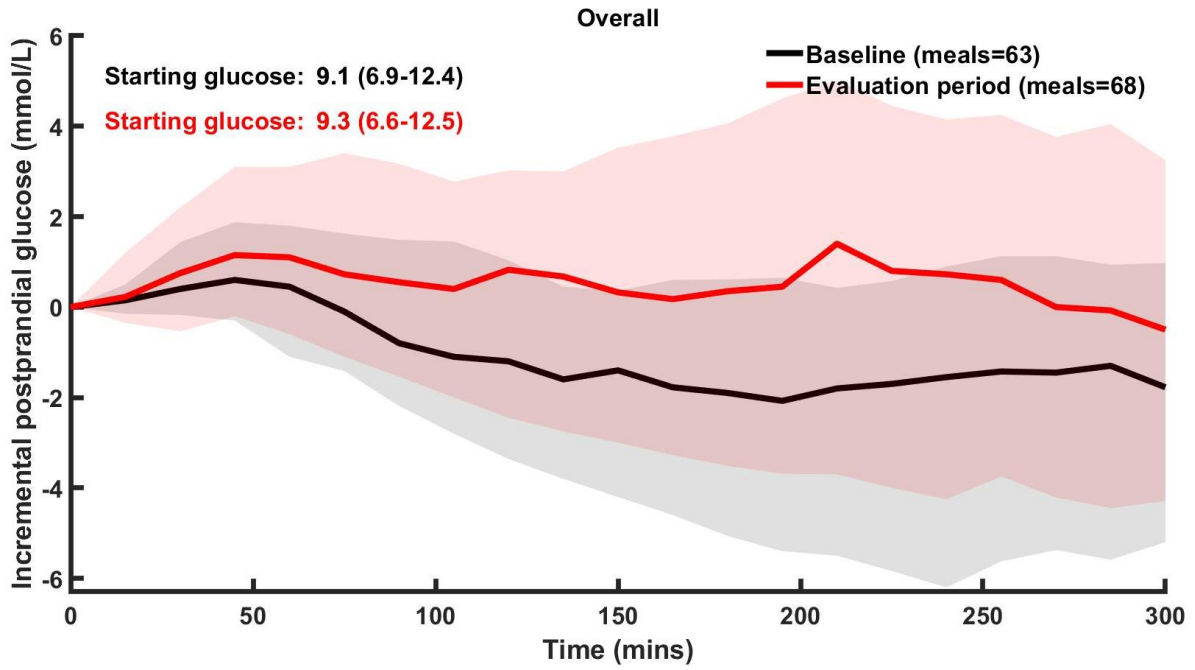


Figure S6. Overall postprandial incremental glucose levels (median and interquartile range) **(a)** after high-fat meals with postprandial exercises and **(b)** non-high-fat meals with postprandial exercises.

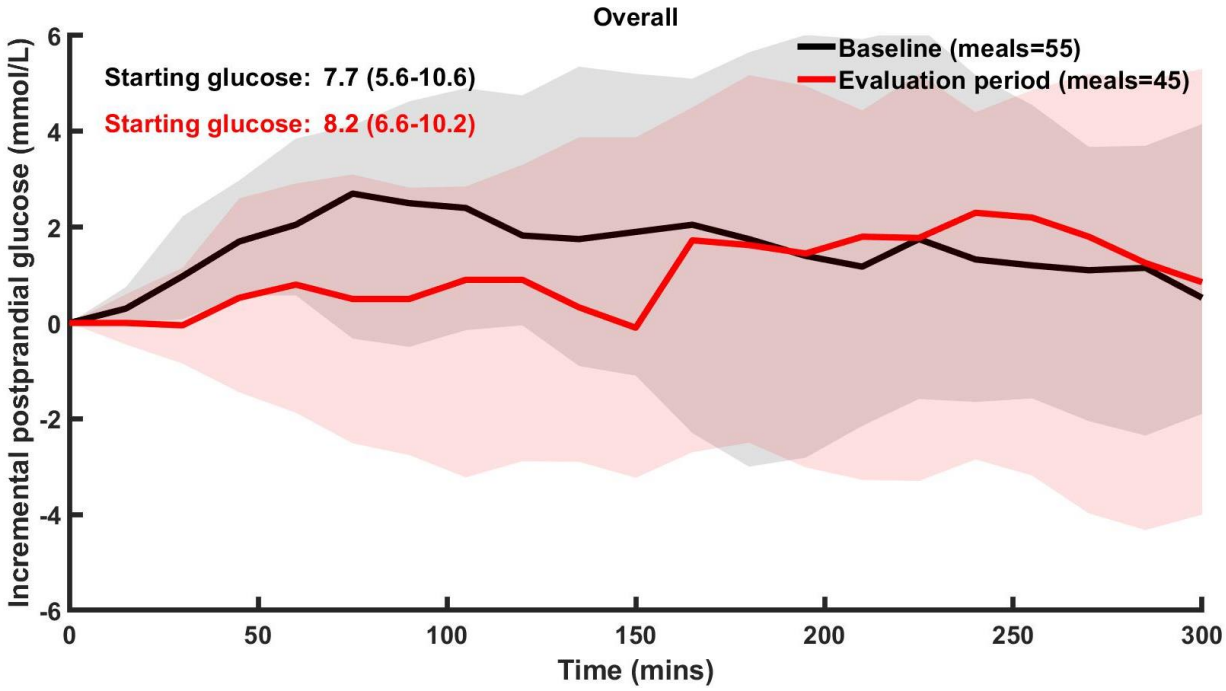


Figure S7. Overall postprandial incremental glucose levels (median and interquartile range) after meals that were preceded, but not followed, by exercise. Source data are provided as a Source Data file.

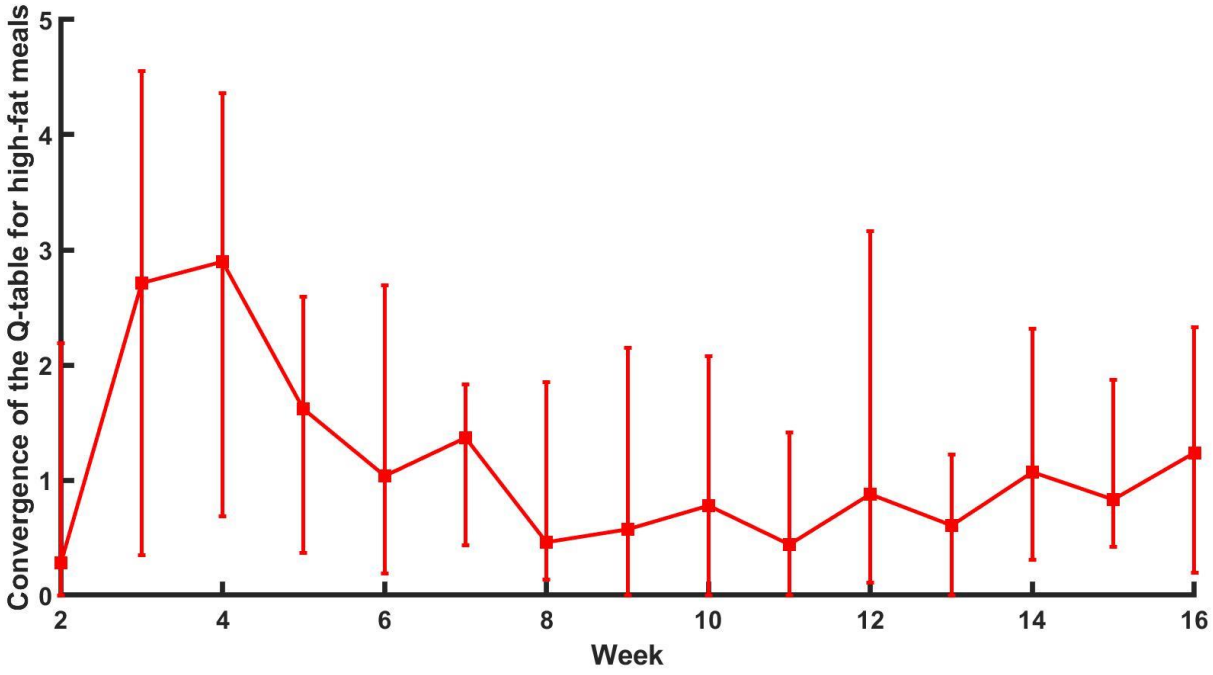


Figure S8. The convergence of the Q-table for the high-fat meals algorithm, defined by the norm (median and quartiles) of the difference between the Q-table for the current week and that of the previous week. Source data are provided as a Source Data file.

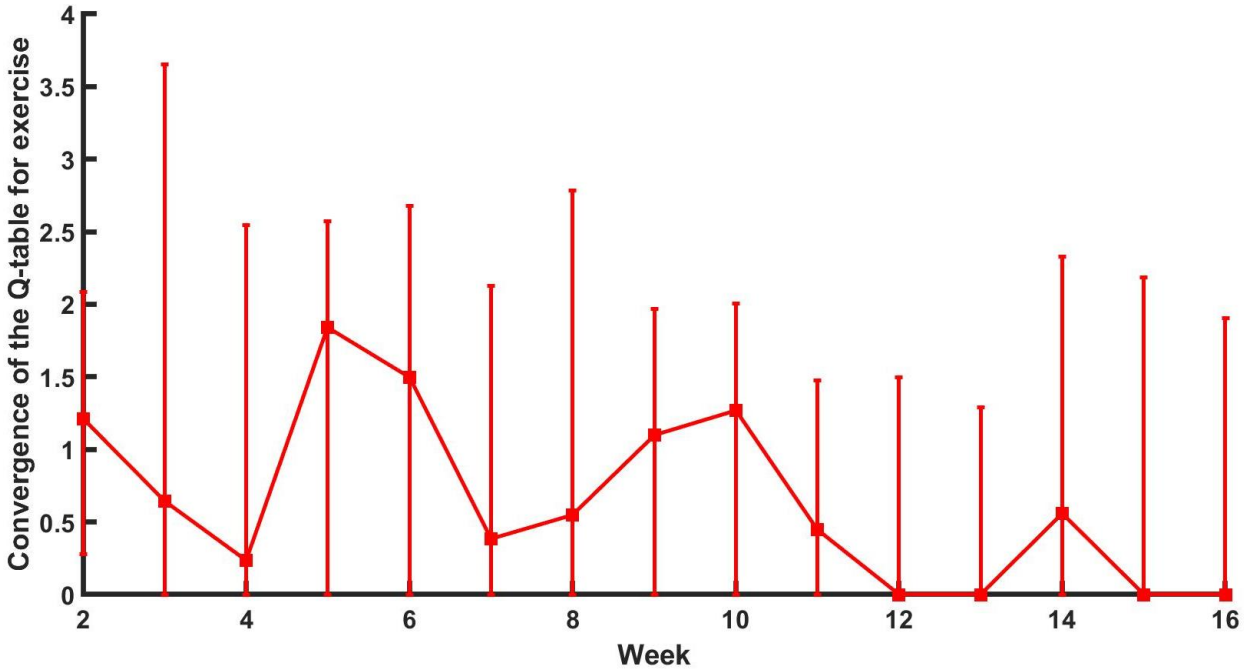


Figure S9. The convergence of the Q-table for the exercise algorithm, defined by the norm (median and quartiles) of the difference between the Q-table for the current week and that of the previous week. Source data are provided as a Source Data file.

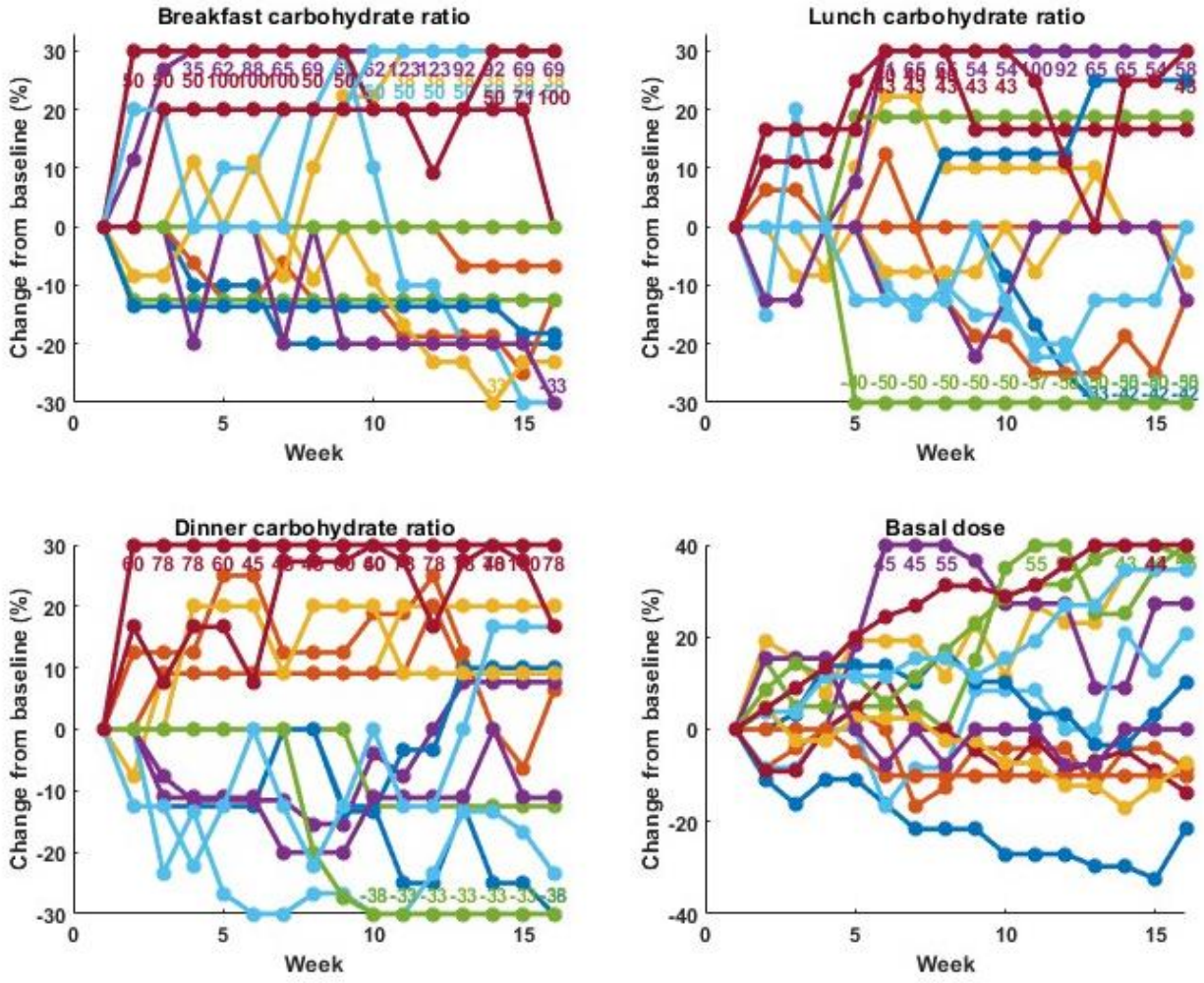


Figure S10. Individual percentage changes in the carbohydrate ratios and the basal dose throughout the study duration (n=14). Source data are provided as a Source Data file.

Table S1. Comparison of the baseline period versus the evaluation period in postprandial outcomes after high-fat meals, while adjusting for effects of age and sex (n=10).

Overall	Baseline period	Evaluation period	P value
Number of High-Fat Meals	111	120	-
iAUC 0-300 min (mmol/L/min)	388±235	52±237	0.04
iAUC 240-300 min (mmol/L/min)	133±55	6±55	0.01
TBR < 3.9 mmol/L	5.0±1.6	2.2±1.6	0.01
Mealtime Insulin (U)	10.5±1.2	9.4±1.2	0.09
Later Bolus Insulin (U)	0.4±0.2	1.4±0.2	<0.001
Correction Bolus Insulin (U)	1.0±0.3	0.5±0.2	0.057

Outcomes are reported as mean ± standard error. Two-side p values are calculated using a linear mixed model. No adjustments were made for multiple comparisons.

Table S2. Comparison of the baseline period versus the evaluation period in postprandial outcomes after meals with postprandial exercises (n=9).

Overall	Baseline period	Evaluation period	P value
After high-fat meals with postprandial exercises			
Number of Exercise Sessions	19	25	-
iAUC 0-300 min (mmol/L/min)	-560±456	128±461	0.06
iAUC 240-300 min (mmol/L/min)	-160±100	0.2±99	0.12
TBR < 3.9 mmol/L	5.2±3.3	3.5±3.2	0.65
Mealtime Insulin (U)	10.0±1.1	7.0±1.7	0.01
Correction Bolus Insulin (U)	0.0±0.3	0.3±0.2	0.40
After non high-fat meals with postprandial exercises			
Number of Exercise Sessions	63	68	-
iAUC 0-300 min (mmol/L/min)	-272±258	158±238	0.052
iAUC 240-300 min (mmol/L/min)	-138±70	-8±64	0.055
TBR < 3.9 mmol/L	5.4±1.1	1.0±1.0	<0.001
Mealtime Insulin (U)	9.5±1.1	8.5±1.2	0.06
Correction Bolus Insulin (U)	0.2±0.1	0.2±0.1	0.60

Outcomes are reported as mean ± standard error. Two-side p values are calculated using a linear mixed model. No adjustments were made for multiple comparisons.

Table S3. Comparison of the baseline period versus the evaluation period in postprandial outcomes after meals that were preceded, but not followed, by exercise (n=9).

Overall	Baseline period	Evaluation period	P value
Meals	55	45	-
iAUC 0-300 min (mmol/L/min)	358±152	89±160	0.20
iAUC 240-300 min (mmol/L/min)	85±43	8±42	0.25
TBR < 3.9 mmol/L	5.8±2.8	4±2.8	0.34
Mealtime Insulin (U)	9.7±1.0	9.4±1.0	0.68
Correction Bolus Insulin (U)	0.0±0.1	0.4±0.2	0.04

Outcomes are reported as mean ± standard error. Two-sided p values are calculated using a linear mixed model. No adjustments were made for multiple comparisons.

Table S4. Comparison of the baseline period versus the evaluation period in postprandial outcomes after postprandial exercises, while adjusting for effects of age and sex (n=9).

Overall	Baseline period	Evaluation period	P value
Number of Exercise Sessions	82	93	-
iAUC 0-300 min (mmol/L/min)	-387±198	134±187	0.008
iAUC 240-300 min (mmol/L/min)	-144±48	-9±44	0.01
TBR < 3.9 mmol/L	5.3±1.2	1.4±1.2	0.002
Mealtime Insulin (U)	9.4±1.2	7.7±1.2	<0.001
Correction Bolus Insulin (U)	0.1±0.1	0.3±0.1	0.32

Outcomes are reported as mean ± standard error. Two-sided p values are calculated using a linear mixed model. No adjustments were made for multiple comparisons.

Table S5. Comparison of glucose and insulin outcomes between the first 10 days and the last 10 days.

	First 10 days (n=14)	Last 10 days (n=14)	Difference, p value
24H OUTCOMES (08:00-08:00)			
Time spent at glucose levels (%):			
3.9–10 mmol/L	48.9±13.6	47.9±13.1	-1.0 (11.6), 0.73
3.9–7.8 mmol/L	30.2±9.8	30.9±9.7	-0.6 (9.3), 0.80
< 2.8 mmol/L	1.8±2.5	1.4±1.7	0.4 (1.9), 0.52
< 3.0 mmol/L	2.4±2.9	1.9±2.2	0.4 (1.9), 0.41
< 3.9 mmol/L	6.3±5.8	6.0±4.1	0.3 (5.4), 0.86
> 7.8 mmol/L	64.5±12.5	64.0±11.9	0.5 (14.0), 0.89
> 10.0 mmol/L	45.0±14.7	46.1±14.9	-1.1 (14.9), 0.75
> 13.9 mmol/L	20.9±14.7	22.0±13.6	-1.0 (12.7), 0.68
> 16.7 mmol/L	10.5±11.1	11.6±9.2	-1.1 (10.6), 0.69
Mean glucose (mmol/L)	10.2±1.9	10.3±1.8	0.1 (1.8), 0.86
SD of glucose (mmol/L)	4.3±1.2	4.5±0.9	-0.2 (0.9), 0.45
CV of glucose (%)	43.0±7.7	44.0±5.4	-1.0 (7.4), 0.40
Basal Insulin (U/day)	29.0±12.0	32.0±17.0	-3.0 (10.1), 0.27
Bolus Insulin (U/day)	25.1±9.9	24.6±7.7	0.47 (5.9), 0.77
Total Insulin (U/day)	54.1±19.0	56.7±17.2	-2.6 (8.9), 0.29
DAY OUTCOMES (08:00-23:00)			
Time spent at glucose levels (%):			
3.9–10.0 mmol/L	48.4±14.7	47.6±14.5	0.7 (13.1), 0.82
3.9–7.8 mmol/L	29.9±10.0	29.2±10.8	0.8 (10.3), 0.76
< 2.8 mmol/L	1.8±2.5	1.4±1.7	0.8 (2.6), 0.26
< 3.3 mmol/L	2.9±4.4	1.8±2.2	1.0 (3.2), 0.23
< 3.9 mmol/L	5.8±6.4	4.4±3.9	1.4 (5.8), 0.38
> 7.8 mmol/L	65.3±13.3	67.3±13.3	-2.0 (14.6), 0.60
> 10.0 mmol/L	45.8±16.0	47.9±16.5	-2.1 (15.6), 0.61
> 13.9 mmol/L	21.7±16.2	23.0±13.6	-1.3 (13.2), 0.64
> 16.7 mmol/L	10.9±12.3	12.3±8.9	-1.4 (10.9), 0.63
Mean glucose (mmol/L)	10.3±2.2	10.5±1.9	-0.2 (1.9), 0.53
SD of glucose (mmol/L)	4.3±1.2	4.5±1.0	-0.2 (1.2), 0.46
CV of glucose (%)	42.1±8.04	43.0±6.7	-0.9 (8.3), 0.62
Bolus Insulin (U/day)	22.3±8.6	19.6±6.8	2.7 (8.6), 0.26
OVERNIGHT OUTCOMES (23:00-08:00)			
Time spent at glucose levels (%):			
3.9–10.0 mmol/L	49.9±15.6	48±17.2	1.9 (17.2), 0.68
3.9–7.8 mmol/L	30.8±12.7	34.1±13.0	-3.2 (12.9), 0.36
< 2.8 mmol/L	1.9±1.9	2.6±3.8	-0.7 (3.0), 0.35
< 3.3 mmol/L	2.9±3.0	3.0±3.2	-0.1 (2.6), 0.80
< 3.9 mmol/L	7.1 (6.3)	8.8±5.5	-1.7 (6.6), 0.34
> 7.8 mmol/L	63.0±15.0	57.8±14.9	5.2 (18.3), 0.31
> 10.0 mmol/L	43.0±16.9	43.1±19.2	-0.19 (21.9), 0.97
> 13.9 mmol/L	18.6±16.2	20.0±17.5	-1.4 (21.9), 0.79
> 16.7 mmol/L	8.9±11.4	9.4±14.0	-0.5 (19.7), 0.91
Mean glucose (mmol/L)	9.9±2.0	9.6±2.1	0.3 (2.8), 0.69
SD of glucose (mmol/L)	4.2±1.2	4.2±1.0	0.0 (1.0), 0.95
CV of glucose (%)	41.9±8.8	43.7±7.5	-1.8 (6.3), 0.23
Bolus Insulin (U/day)	2.9±3.4	5.0±6.8	-2.1 (5.5), 0.18
HbA1c (%)	8.5±1.1	8.2±1.0	-0.3±0.9, 0.12

Outcomes are reported as mean ± standard deviation. Two-sided p values are calculated using a paired t test. No adjustments were made for multiple comparisons.

Table S6. Additional comparisons of the baseline period versus the evaluation period in postprandial outcomes after high-fat meals with no postprandial exercises (n=10).

	Baseline period	Evaluation period	P value
Number of High-Fat Meals	111	120	-
Time spent at glucose levels (%):			
3.9–10 mmol/L	41.2±4.2	45.9±4.3	0.31
3.9–7.8 mmol/L	23.3±3.2	24.6±3.1	0.69
< 2.8 mmol/L	0.9±0.4	0.4±0.4	0.23
< 3.0 mmol/L	1.4±0.5	0.5±0.4	0.12
> 7.8 mmol/L	71.6±3.9	72.4±3.8	0.84
> 13.9 mmol/L	25.1±4.9	24.2±4.9	0.80
> 16.7 mmol/L	14.3±3.9	12.2±3.8	0.50

Outcomes are reported as mean ± standard error. Two-sided p values are calculated using a linear mixed model. No adjustments were made for multiple comparisons.

Table S7. Additional comparisons of the baseline period versus the evaluation period in postprandial outcomes with exercises (n=9).

	Baseline period	Evaluation period	P value
Number of Exercise Sessions	82	93	-
Time spent at glucose levels (%):			
3.9–10 mmol/L	50.3±6.5	48.9±6.1	0.62
3.9–7.8 mmol/L	35.0±4.0	24.4±3.7	0.04
< 2.8 mmol/L	0.8±0.5	0.3±0.4	0.50
< 3.0 mmol/L	0.8±0.5	0.3±0.4	0.49
> 7.8 mmol/L	62.2±4.2	74.1±3.9	0.03
> 13.9 mmol/L	20.5±5.5	24.3±5.2	0.40
> 16.7 mmol/L	10.5±4.4	13.9±4.2	0.36

Outcomes are reported as mean ± standard error. Two-sided p values are calculated using a linear mixed model. No adjustments were made for multiple comparisons.

Table S8. Comparison of the survey outcomes in the admission visit versus the study end visit (n=14).

Survey	Admission Visit	Study End Visit	P value
HFS-II	1.9±0.5	2.1±0.8	0.72
DTSQ	4.0±0.9	4.6±1.1	0.22
MAUQ	-	5.4±0.9	-

HFS-II means Hypoglycemia Fear Survey-II scale, DTSQ means Diabetes Treatment Satisfaction Questionnaire, and MAUQ means Mobile Application Usability Questionnaire. There was a total of 18 questions in the HFS-II, each with scores ranging from 1 to 5, where 1 indicates the best and 5 indicates the worst. There was a total of 9 questions in the DTSQ survey, each with scores ranging from 0 to 6, where 0 indicates the worst and 6 indicates the best. There was a total of 16 questions in the MAUQ survey, each with scores ranging from 0 to 6, where 0 indicates the worst and 6 indicates the best. The MAUQ survey was conducted only at the study end visit. Outcomes are reported as mean ± standard deviation. Two-sided p-values were calculated using a paired t-test. No adjustments were made for multiple comparisons.

Table S9. Total number of hypoglycemia events in the baseline and evaluation periods (n=9).

	Total Exercises		**Total hypoglycemia events (%)**	
Participant ID	Baseline period	Evaluation period	Baseline period	Evaluation period
205	8	8	50%	0%
206	2	8	100%	25%
207	9	10	11%	17%
208	5	18	67%	36%
209	8	12	50%	8%
210	1	3	0%	0%
211	24	7	12%	0%
212	12	10	37%	7%
213	12	13	33%	7%
Mean± standard deviation	9.0±6.8	9.9±4.2	40.0±31.2%	11.1±12.5%

The hypoglycemia event per exercise is defined as a glucose value less than 3.9 and is limited to a maximum of one event per exercise session.

Participant feedback on system functionality

Analysis of the semi-structured interviews using inductive thematic analysis revealed informative feedback that could guide future developments of the iBolusV2 app. Some users felt that the need to announce the exercise up to four hours in advance is not always practical, which motivates the introduction of a future feature to announce exercise at any time:

> *"Sometimes I didn't plan for hours ahead. So sometimes I would only know, one or two hours in advance." (Participant 15)*

> *"Unless you have a real schedule that you're going to follow, it's a little bit less practical." (Participant 9)*

Users also appreciated the ability to announce the time of the exercise (early vs late postprandial) and having the insulin dose recommendations adjusted accordingly:

> *"It definitely helped first of all having the within 90 minutes and after." (Participant 13)*

> *"If I was doing exercise within 90 minutes obviously it would give me a little bit less-lesser of an insulin bolus than it would if I wasn't…so I did it and it worked out." (Participant 10)*

Several participants would have liked to be able to log snacks in the app:

> *"I think what I most would have liked to see is somewhere where I can log what I'm eating that I'm not getting insulin for." (Participant 8)*

> *"It would have been nice to be able to add like a snack, you know, like if I was eating something, there was nowhere to log it." (Participant 9)*

> *"If there was like a feature for me to enter instead of a meal if I'm having just like a small snack, because sometimes I would only have like a bowl of chips, or maybe like a granola bar or something, just a quick snack, to maybe log that instead of like a meal" (Participant 15)*

Participants also appreciated the reminder to deliver the second bolus to manage high-fat meals:

> *"It helps because it reminds you to check again in two hours. So that was great." (Participant 3)*

"In the past, when I had like, let's say pizza, um, I would calculate just for the meal and take that amount, um I would drop into hypo and then immediately spike into hyper and stay really high for a long time. So that part of the app was really useful and helpful." (Participant 6)

Details of Reinforcement Learning Algorithms

Problem formulation

The reinforcement learning problem can be modeled as a discrete Markov decision process $(\mathbf{S}, \mathbf{A}, \mathbf{P}, r, \gamma)$, with state space $\mathbf{S}$ (physiological state), action space $\mathbf{A}$ (insulin adjustments), transition dynamics $\mathbf{P}(\mathbf{s_{k+1}}|\mathbf{s_k}, r_k)$, reward r (improvement in glycemic outcomes), and a discount factor γ.[1] The agent in the environment receives a reward $r_k(\mathbf{s_k}, a_k, \mathbf{s_{k+1}}) \epsilon \mathbb{R}$ by taking an action a_k in a state $\mathbf{s_k}$ and reaching at a state $\mathbf{s_{k+1}}$. The goal of the agent is to maximize the accumulated reward:

$$R_k = \sum_{i=k}^{\infty} \gamma^{i-k} r_i(\mathbf{s_i}, a_i, \mathbf{s_{i+1}}) \tag{1}$$

The agent selects its actions based on a policy $\pi: \mathbf{S} \rightarrow \mathbf{A}$. The value-function $J_k^{\pi}: \mathbf{S}x\mathbf{A} \rightarrow \mathbb{R}$ under the policy π can be defined as the expected return in the future time interval T when the policy π is followed:

$$J_k^{\pi}(\mathbf{s_k}) = \mathbb{E}_{\pi}\left[\sum_{i=k}^{k+T} \gamma^{i-k} r_i | \mathbf{s} = \mathbf{s_k}\right] \tag{2}$$

The optimal policy at time-step k is defined as $\pi^*(\mathbf{s_k}) = \arg\max_{\pi} J_k^*(\mathbf{s_k})$. Alternatively, we can find the optimal policy π^* for the state-action pair $(\mathbf{s_k}, a_k)$ using the optimal action-value function:[1]

$$\mathbf{Q_k^*}(\mathbf{s_k}, a_k) = \sum_{s_{k+1} \epsilon S} p_{\mathbf{s_k s_{k+1}}}^{a_k} [r_k + \gamma \max_{a_u} J_k^*(\mathbf{s_{k+1}})] \tag{3}$$

Using the temporal difference method,[1] $\mathbf{Q_k^*}(\mathbf{s_k}, a_k)$ is estimated by recursively solving the Bellman optimality equation:

$$\mathbf{Q_{k+1}}(\mathbf{s_k}, a_k) = \mathbf{Q_k}(\mathbf{s_k}, a_k) + \alpha[r_k + \gamma \max_{a_u} \mathbf{Q_k}(\mathbf{s_{k+1}}, a_u) - \mathbf{Q_k}(\mathbf{s_k}, a_k)] \tag{4}$$

In our problem, we aim to learn optimal insulin dosing for high-fat meals per each main meal (breakfast, lunch, and dinner) and optimal pre-meal insulin adjustments for postprandial aerobic exercises per each main meal. Using multiple independent agents (one for each meal) may be inefficient due to data scarcity per each main meal in real-world settings. Therefore, we adopt a collaborative multi-agent reinforcement learning approach to leverage agents' learning

experiences between different meals (breakfast, lunch, and dinner), to facilitate faster convergence. More specifically, we assume that each agent j receives its own reward r_k^j based on its individual action a_k^j in a state $\mathbf{s_k^j}$ using a global Q-value function (weighted average Q-values of all agents).[2-4] The environment will consist of three agents for high-fat meals (one for breakfast, lunch, and dinner) and six agents for postprandial aerobic exercises (2 per each meal; one for exercises within 90 minutes of each meal and another for later than 90 minutes). At any period (e.g., breakfast), there will be one agent (agent j) that interacts with the environment using the action policy evaluated by the global Q-value function. Then, the Q-value of agent j is updated as follows:

$$\mathbf{Q_{k+1}^j}(\mathbf{s_k^j}, a_k^j) = \mathbf{Q_k^j}(\mathbf{s_k^j}, a_k^j) + \alpha[r_{k+1}^j + \frac{\gamma}{N} \max_{a_u^j} \mathbf{Q_k^G}(\mathbf{s_{k+1}^j}, a_u^j) - \mathbf{Q_k^j}(\mathbf{s_k^j}, a_k^j)] \quad (5)$$

where $\boldsymbol{N}$ equals the number of agents. The global Q-value $\mathbf{Q_k^G}(\mathbf{s_k^j}, a_u^j)$ is the weighted sum of all agents' Q-values:

$$\mathbf{Q_k^G}(\mathbf{s_k^j}, a_k^j) = \sum_{j=1}^{N} w^j(\mathbf{s_k^j}, a_k^j)\mathbf{Q_k^j}(\mathbf{s_k^j}, a_k^j) \quad (6)$$

where the weight w^j for each agent j is defined as:

$$w^j(\mathbf{s_k^j}, a_k^j) = \begin{cases} \frac{\text{Count}^j(\mathbf{s_k^j}, a_k^j)}{\sum_{j=1}^{N} \text{Count}^j(\mathbf{s_k^j}, a_k^j)}, \forall j \in N & \text{if } \text{count}^c(\mathbf{s_k^j}, a_k^j) \leq 4 \\ \left(1 + \text{temp}^j(\mathbf{s_k^j}, a_k^j)\right) \frac{\text{Count}^j(\mathbf{s_k^j}, a_k^j)}{\sum_{j=1}^{N} \text{Count}^j(\mathbf{s_k^j}, a_k^j)} & \text{if } \text{count}^c(\mathbf{s_k^j}, a_k^j) > 4 \text{ and } \leq 8, \forall j = c \\ \rho(\mathbf{s_k^j}, a_k^j)\left(2 - \text{temp}^c(\mathbf{s_k^c}, a_k^c)\right) \frac{\text{Count}^j(\mathbf{s_k^j}, a_k^j)}{\sum_{j=1}^{N} \text{Count}^j(\mathbf{s_k^j}, a_k^j)} & \text{if } \text{count}^c(\mathbf{s_k^j}, a_k^j) > 4 \text{ and } \leq 8, \forall j \neq c \\ 1 \; \forall j = \text{c and } 0 \; \forall j \neq \text{c} & \text{else } \text{count}^c(\mathbf{s_k^j}, a_k^j) > 8 \end{cases} \quad (7)$$

where c is the current agent, $\text{count}^j(\mathbf{s_k^j}, a_k^j)$ is a counter variable that counts how many times the agent j visited the state-action pair $(\mathbf{s_k^j}, a_k^j)$, $\text{temp}^j(\mathbf{s_k^j}, a_k^j)$ is a counter variable that increments by 0.2 when the agent j visits the state-action pair $(\mathbf{s_k^j}, a_k^j)$, and $\rho(\mathbf{s_k^j}, a_k^j)$ is the relative experience of agent j at the state-action pair $(\mathbf{s_k^j}, a_k^j)$ calculated by dividing the total number of visits of agent j by the total number of visits by other agents to the state-action pair $(\mathbf{s_k^j}, a_k^j)$ as follow:

$$\rho(\mathbf{s_k^j}, a_k^j) = \frac{\text{Count}^{\text{j}}\left(\mathbf{s_k^j}, a_k^j\right)}{\sum_{j=1}^{N} \text{Count}^{\text{j}}\left(\mathbf{s_k^j}, a_k^j\right) - \text{Count}^{\text{c}}\left(\mathbf{s_k^c}, a_k^c\right)} \quad (8)$$

In our multi-agent reinforcement learning, each agent learns not only from its own experiences but also benefits from the collective knowledge gained by other agents (Equation. 6). This collaborative learning strategy could mitigate against the scarcity of data.

Our multi-agent reinforcement learning algorithm bears resemblance to the coordinated reinforcement learning approaches[5-7] where the global Q-function (Equation 6), being a target for the algorithm, is calculated using the weighted sum of local Q-functions. The main difference between these approaches[5-7] and ours is that they calculate the global Q-function using equal weights from all agents in contrast to our method where we weight each agent based on its visits to a specific state-action pair divided by the total visits of all agents to that same state-action pair.

We will now discuss the algorithms for high-fat meals, exercise, and basal doses.

Multi-agent Q-learning method for high-fat meals

In this sub-section, we define the states, actions, and rewards of the multi-agent Q-learning algorithm.

State space

We selected features in the algorithm's state vector for high-fat meals to steer the late postprandial glucose levels into the target range while avoiding hypoglycemia. Additionally, we included a feature comprising glucose levels in the early and late postprandial periods, considering the impact of high-fat meals on both time periods. Therefore, the state vector $\mathbf{s_k^j}$ of each agent *j* for each main meal (breakfast, lunch, and dinner) is composed of (i) the 5-hour postprandial percentage time spent with glucose levels below 3.9 mmol/L, $\text{Hypo}_{\text{k}}^{\text{j}}$, (ii) the ratio of postprandial incremental area under the glucose levels curve (iAUC) in the period 0 and 2-hour to the period 2 and 5-hour after the meal, $R_k^j = \frac{\text{0\_2h\_iAUC}_{\text{k}}^{\text{j}}}{\text{2\_5h\_iAUC}_{\text{k}}^{\text{j}}}$, and (iii) the postprandial error in glucose levels defined as:

$$E_k^j = (G_{min}(k) - G_T) \quad (9)$$

where $G_{min}(k)$ is the minimum glucose level in the period 3-6 hours after the meal.

For each agent j, the discretized state vector is defined as:

$$\begin{aligned}\mathbf{S^j} &= \{\mathbf{s^j}|\mathbf{s}_{mno} \\ &= \left(\mathrm{Hypo}^{\mathrm{j}}_{\mathrm{k,m}}, R^j_{k,n}, E^j_{k,o}\right), m\epsilon[1,2,\dots,M], n\epsilon[1,2,\dots,N], o\epsilon[1,2,\dots,O]\}\end{aligned} \tag{10}$$

where N, M, and O represent the number of discrete segments of $\mathrm{Hypo}^{\mathrm{j}}_{\mathrm{k}}$, R^j_k, and E^j_k, respectively. The choice of *N, M,* and *O* is a trade-off between the learning period and the accuracy. A discrete state space was chosen to accommodate the scarcity of the training and validation data which renders continuous function approximators like neural networks impracticable to use. We define the goal state for the agent j, $\mathbf{s^j_g}$, as $\mathrm{Hypo}^{\mathrm{j}}_{\mathrm{k}} = 0\%$, $\left|R^j_k\right| \geq 0.6$, and $\left|E^j_k\right| \leq 1.2$ mmol/L.

Action space

Let $\mathbf{A^j}(\mathbf{s^j})$ represents the set of possible actions at state $\mathbf{s^j}$ defined as:

$$\mathbf{A^j}(\mathbf{s^j}) = \{a^j|1,2,3,4,5,6\} \tag{11}$$

The action space was chosen to be in the discrete domain to reduce the computational burden of running the algorithm. The mapping from agents' actions to $F^j_{u,k+1}$ and $F^j_{L,k+1}$ changes are as follows:

$$F^j_{u,k+1} = F^j_{u,k} + f_1(a^j)\, k_1\, F^j_{u,k} \tag{12}$$

$$F^j_{L,k+1} = F^j_{L,k} + f_2(a^j)\, k_2\, F^j_{L,k} \tag{13}$$

where $f_1(a^j)$ and $f_2(a^j)$ equal to $1, -1$ or 0, denote increasing, decreasing, and unchanging F^j_u and F^j_L values relative to previous day's values, respectively, and $k_1 = \left|0.05\left|E^j_k\right| - 0.05\right|$ and $k_2 = \frac{0.2}{\max\left(\left|R^j_k\right|, 0.2\right)}$ are gain parameters. The mapping from the agent's action a^j to $f_1(a^j)$ and $f_2(a^j)$ is given in Table S5.

Table S5. Mapping from agent's action to the qualitative changes in fat ratios

Action a^j	Qualitative changes in upfront and later bolus ratios by $f_1(a^j)$ and $f_2(a^j)$
1	$f_1(a^j) = 0$ and $f_2(a^j) = 0$
2	$f_1(a^j) = -1$ and $f_2(a^j) = -1$
3	$f_1(a^j) = -1$ and $f_2(a^j) = 0$
4	$f_1(a^j) = 0$ and $f_2(a^j) = 1$
5	$f_1(a^j) = 1$ and $f_2(a^j) = 0$
6	$f_1(a^j) = 0$ and $f_2(a^j) = -1$

Moreover, the algorithm modifies the later bolus based on the glucose value at 2 hours after the meal G_{k+2h}^{j} as follows:

$$L_B^j = \begin{cases} 0 & \text{if } G_{k+2h}^{j} < 6 \text{ mmol/L} \\ 0.5L_B^j & \text{if } 6 \text{ mmol/L} \le G_{k+2h}^{j} \le 7 \text{ mmol/L} \\ L_B^j & \text{otherwise} \end{cases} \tag{14}$$

Reward function

After taking an action a_k^j, the agent receives a reward r_{k+1}^j as follows:

$$r_{k+1}^j = \begin{cases} 10e^{\frac{\left|\Delta E_k^j\right|}{10}} & \text{if } \left(\left|E_{k+1}^j\right| < \left|E_k^j\right|\right) \,\&\, \left(\text{Hypo}_{k+1}^{j} < \text{Hypo}_{k}^{j}\right) \,\&\, \left(\left|R_{k+1}^j\right| \ge \left|R_k^j\right|\right) \\ 10 & \text{if } (\mathbf{s_{k+1}^j} = \mathbf{s_g^j}) \\ 1 & \text{if } (\mathbf{s_{k+1}^j} = \mathbf{s_k^j}) \,\&\, (\mathbf{s_k^j} \neq \mathbf{s_g^j}) \\ -10ne^{\frac{\left|\Delta E_k^j\right|}{10}} & \text{otherwise} \end{cases} \tag{15}$$

where n equals to 1 if $\text{Hypo}_{k}^{j} > 0$ and equals to 2 otherwise.

The agent receives a positive reward for its action if $\left|E_{k+1}^j\right|$ and Hypo_{k+1}^{j} are dampened, and R_{k+1}^j is increased. If the action led to the goal state $\mathbf{s_g^j}$, then the agent receives a large positive reward. If the action did not lead to change in the state and $\text{Hypo}_{k}^{j} = 0$, the agent receives a small positive reward. In all other cases, the agent receives a negative reward.

The global Q-value in the multi-agent learning algorithm for high-fat meals is calculated from (6) with N = 3. The detail of the multi-agent learning algorithm for high-fat meals is given in Algorithm 1.

Algorithm 1: Multi-agent Q-learning algorithms

1. **Initialize** $\gamma = 0.5$, $\alpha = 0.55$, $\varepsilon = 0.1$.
2. **Initialize** $\mathbf{Q^z}(\mathbf{s_k^z}, a_k^z)$, $\forall \mathbf{s^z} \epsilon \mathbf{S^z}$, $\forall a^z \epsilon \mathbf{A^z}$ using common clinical guidelines, where $z = j$ denotes the agent for high-fat meals and $z = t$ denotes the agent for postprandial aerobic exercises. We initialize $\mathbf{Q^z}(\mathbf{s_k^z}, a_k^z)$ such that higher values are assigned to better state-action pairs, an approach that bears similarity to the concept of reward shaping to help the agent learn optimal behavior.[8]
3. **Evaluate** current state $\mathbf{s_k^z}$.
4. **Repeat**

a. Choose action a_k^z from $\boldsymbol{s_k^z}$ using ε-greedy policy.[1]

b. From an action a_k^z, obtain (E_{w90}^w and E_{L90}^l) or (F_u^j and F_L^j) using equations (12-13) or (19-20).

c. Apply $\left(E_{w90}^w \text{ and } E_{L90}^l\right)$or $\left(F_u^j \text{ and } F_L^j\right)$, observe the reward r_{k+1}^z and the next state $\mathbf{s_{k+1}^z}$.

d. Update the action value function $\mathbf{Q_{k+1}^z}(\mathbf{s_k^z}, a_k^z) \leftarrow \mathbf{Q_k^z}(\mathbf{s_k^z}, a_k^z) + \alpha[r_{k+1}^z + \frac{\gamma}{n} \max_{a_u^j} \boldsymbol{Q_k^G}(\mathbf{s_{k+1}^z}, a_u^z) - \boldsymbol{Q_k^z}(\mathbf{s_k^z}, a_k^z)]$.

e. $\mathbf{s_k^z} \leftarrow \mathbf{s_{k+1}^z}$.

Multi-agent Q-learning method for exercise management

In this sub-section, the states, actions, and rewards for the multi-agent Q-learning algorithm are defined. The rationale behind choosing discrete state space, discrete action space, and the exponential form in the reward function is the same as for the algorithm of high-fat meals described above.

State space

The state vector $\mathbf{s_k^w}$ or $\mathbf{s_k^l}$ of each agent w or l (for postprandial aerobic exercise happening within 90 minutes or later than 90 minutes of mealtime) is composed of (i) the 5-hour postprandial percentage time spent with glucose levels below 3.9 mmol/L (hypoglycemia), $\mathrm{eHypo_k^t}$, (ii) the postprandial error in glucose levels, E_k^t, similar to (9), (iii) hypoglycemia treatments in the period 5-hour after meal, $\mathrm{htreat_k^t}$, and (iv) the rate of change of glucose levels in the 1-hour period following the minimum glucose level in the period 1 to 5-hour after meal, $\mathrm{ROC_k^t}$.

For each agent t, the discretized state vector is defined as:

$$\begin{aligned}\mathbf{S^t} = \{\mathbf{s^t}|\mathbf{s}_{pqrs} \\ = \left(\mathrm{eHypo_{k,p}^t}, E_{k,q}^t, \mathrm{htreat_{k,r}^t}, \mathrm{ROC_{k,s}^t}\right), p\epsilon[1,2,\dots,P], q\epsilon[1,2,\dots,Q], r\epsilon[1,2,\dots,R], \\ s\epsilon[1,2,\dots,S]\}\end{aligned} \quad (16)$$

where P, Q, R, and S represents the number of discrete segments of $\mathrm{Hypo_k^t}$, E_k^t, $\mathrm{htreat_k^t}$ and $\mathrm{ROC_k^t}$, respectively. The superscripts $t = w$ and $t = l$ denotes the learning agents for postprandial aerobic exercises happening within and later than 90 minutes after meal, respectively. We define

the goal state for the agent w, $\mathbf{s_g^t}$, as $\text{eHypo}_\text{k}^\text{t} = 0, |E_k^t| \leq 1.2$ mmol/L, $\text{htreat}_\text{k}^\text{t}$ =0, and $\left|\text{ROC}_\text{k}^\text{t}\right| \leq 0.02 \frac{mmol}{L\,min}$.

Action space

Let $\mathbf{A^w}(\mathbf{s^w})$ and $\mathbf{A^l}(\mathbf{s^l})$ represent the set of possible actions at states $\mathbf{s^w}$ and $\mathbf{s^l}$, respectively**.** For all states, we have chosen the action space $\mathbf{A^w}$ and $\mathbf{A^l}$ as follows:

$$\mathbf{A^w}(\mathbf{s^w}) = \{a^w|1,-1,0\} \quad (17)$$

$$\mathbf{A^l}(\mathbf{s^l}) = \{a^l|1,-1,0\} \quad (18)$$

where $1, -1, 0$ denote increasing, decreasing, and unchanging, respectively, of E_{W90}^w and E_{L90}^l values relative to previous day's values. The mappings from the agents' actions to $E_{W90,k+1}^w$and $E_{L90,k+1}^l$ changes are as follows:

$$E_{W90,k+1}^w = E_{W90,k}^w + a^w\, k_e\, E_{W90,k}^w \quad (19)$$

$$E_{L90,k+1}^l = E_{L90,k}^l + a^l\, k_e E_{L90,k}^l \quad (20)$$

where a^w and a^l equal to $1, -1$ or 0, denote increasing, decreasing, and unchanging E_{W90}^w and E_{L90}^w values relative to previous day's values, respectively, and $k_e = |0.05|E_k^t| - 0.05|$.

Reward Function

After taking an action a_k^t, the agent, the next day, receives a reward r_{k+1}^t as follows:

$$r_{k+1}^t = \begin{cases} 10e^{\frac{|\Delta E_k^t|}{10}} & \text{if } (\text{eHypo}_{\text{k+1}}^\text{t} < \text{eHypo}_\text{k}^\text{t}) \;\&\; (|E_{k+1}^t| \leq |E_k^t|) \;\&\; (\text{ROC}_{\text{k+1}}^\text{t} \leq \text{ROC}_\text{k}^\text{t}) \\ 1 & \text{if } (\mathbf{s_{k+1}^t} = \mathbf{s_k^t}) \\ 100 & \text{if } (\mathbf{s_{k+1}^t} = \mathbf{s_g^t}) \\ -10ne^{\frac{|\Delta E_k^t|}{10}} & \text{otherwise} \end{cases} \quad (21)$$

where n equals to 1 if $\text{eHypo}_\text{k}^\text{t} > 0$ and equals to 2 otherwise.

The agent receives a positive reward for its action if $\text{eHypo}_{\text{k+1}}^\text{t}$, $|E_{k+1}^t|$, and $\text{ROC}_{\text{k+1}}^\text{t}$ are dampened. If the action led to the goal state $\mathbf{s_g^t}$, then the agent receives a large positive reward. If the action did not lead to a change in the state and $\text{eHypo}_\text{k}^\text{t} = 0$, the agent receives a small positive reward. In all other cases, the agent receives a negative reward.

The global Q-value in the multi-agent learning algorithm for postprandial aerobic exercise managemt with N = 3 is calculated as follows:

$$\mathbf{Q^G}(\mathbf{s_k^t}, a_k^t) = \theta_1 \sum_{w=1}^{N} w^w(\mathbf{s_k^w}, a_k^w)\mathbf{Q}^w(\mathbf{s_k^w}, a_k^w) + \theta_2 \sum_{l=1}^{N} w^l\left(\mathbf{s_k^l}, a_k^l\right)\mathbf{Q^l}\left(\mathbf{s_k^l}, a_k^l\right) \quad (22)$$

where w^w and w^l are obtained from (8) for $j = w$ and $j = l$, and θ_1 and θ_2 are defined as:

$$[\theta_1\ \theta_2] = [\frac{\text{count}^{\text{w}}(\mathbf{s_k^w}, a_k^w)}{\text{count}^{\text{w}}(\mathbf{s_k^w}, a_k^w) + \text{count}^{\text{l}}\left(\mathbf{s_k^l}, a_k^l\right)}\ \frac{\text{count}^{\text{l}}\left(\mathbf{s_k^l}, a_k^l\right)}{\text{count}^{\text{w}}(\mathbf{s_k^w}, a_k^w) + \text{count}^{\text{l}}\left(\mathbf{s_k^l}, a_k^l\right)}] \quad (23)$$

Learning algorithm for basal

In this sub-section, the states, actions, and rewards are defined for the basal learning algorithm. The rationale behind choosing discrete state space, discrete action space, and the exponential form in the reward function is the same as that for the high-fat meals and exercise algorithms described above.

State space

The features in the state vector for the basal learning algorithm were chosen to steer the nighttime glucose level to the target range while avoiding hypoglycemia. The state $\mathbf{s_k^b}$ consists of (i) the mean glucose error, ΔG_k, calculated as the mean glucose G_k^{mean} between 2am and 7am minus the target glucose level G_T, (ii) the percentages time spent with glucose levels below 3.9 and above 10.0 mmol/L, T_k^{hypo} and T_k^{hyper}, respectively, in the period between 2am and 7am, (iii) the rate of change of glucose in the period between 4am and 7am, ROC_k, and (iv) hypoglycemia treatments in the period between 10pm and 2am, htreat_k.

The discretized state vector is defined as:

$$\begin{aligned}\mathbf{S} &= \{\mathbf{s^b}|\mathbf{s}_{abcde} \\ &= \left(\Delta G_{k,a}, T_{k,b}^{hypo}, T_{k,c}^{hyper}, \text{ROC}_{\text{k,d}}, \text{htreat}_{\text{k,e}}\right), a\epsilon[1,2,\dots,A], b\epsilon[1,2,\dots,B], c\epsilon[1,2,\dots,C], \\ &\qquad d\epsilon[1,2,\dots,D], e\epsilon[1,2,\dots,E]\}\end{aligned} \quad (24)$$

where A, B, C, D, and F represents the number of discrete segments of $\Delta G_k, T_k^{hypo}, T_k^{hyper}$, $\text{ROC}_\text{k}^\text{t}$, and htreat_k, respectively. We define the goal state $\mathbf{s_g^b}$ as $|\Delta G_k| \leq 1$ mmol/L, $T_k^{hypo} = 0$, $T_k^{hyper} = 0$, $|\text{ROC}_\text{k}| \leq 0.02\frac{mmol}{L\ min}$, and $\text{htreat}_\text{k} = 0$.

Action Space

Let $\mathbf{A}\left(\mathbf{s^b}\right)$ be the set of all possible actions at state $\mathbf{s^b}$. For all states, we have defined the action

space A as follows:

$$\mathbf{A}(\mathbf{s^b}) = \{a^b|1, -1, 0\} \tag{25}$$

where $1, -1, 0$ represent increasing, decreasing, and not changing the basal dose relative to previous day's values, respectively.

Reward function

After taking an action a_k^b, the agent receives a reward r_k^b as follows,

$$r_k^b = \begin{cases} 10e^{\frac{|\Delta G_{k+1}|}{10}} & \text{if } \left(T_{k+1}^{\text{hyper}} \leq T_k^{\text{hyper}}\right) \,\&\, (\Delta G_{k+1} \leq \Delta G_k) \,\&\, \left(T_{k+1}^{\text{hypo}} < T_k^{\text{hypo}}\right) \\ 50 & \text{if } \left(\mathbf{s_{k+1}^b} = \mathbf{s_g^b}\right) \\ 1 & \text{if } \left(\mathbf{s_{k+1}^b} = \mathbf{s_k^b}\right) \\ -10ne^{\frac{|\Delta G_{k+1}|}{10}} & \text{otherwise} \end{cases} \tag{26}$$

where n is a scaling multiplier equals to 1 if glucose levels do not fall below 4.0 mmol/L in the period between 2am and 7am and equals to 2 otherwise.

The the agent receives a positive reward for its action if ΔG_k, T_k^{hypo}, and T_k^{hyper} are dampened. If the selected action led to the goal state $\mathbf{s_g^b}$, then the agent receives a large positive reward. If the selected action did not lead to change in the state, then the agent receives a constant reward of 1. In all other cases, the agent receives a negative reward. The details of the learning algorithm for basal estimation are given in **Algorithm 2**.

Algorithm 2: Learning algorithm for basal

1. **Initialize** $\gamma = 0.9$, $\alpha = 0.55$, $\varepsilon = 0.1$.
2. **Initialize** $\mathbf{Q}(\mathbf{s^b}, a^b)$, $\forall \mathbf{s^b} \epsilon \mathbf{S^b}$, $\forall a^b \epsilon \mathbf{A}$ using clinical guidelines. The rationale of initializing $\mathbf{Q}(\mathbf{s^b}, a^b)$ using clinical guidelines is as explained in **Algorithm 1**.
3. **Evaluate** current state $\mathbf{s_k^b}$.
4. **Repeat**
   a. Choose action a_k^b from $\mathbf{s_k^b}$ using ε-greedy policy.[1]
   b. From an action a_k^b, obtain B_{k+1} using the following:

$$B_{k+1} = \begin{cases} B_k - k_1 B_k & \text{if } T_k^{\text{hypo}} > 5\% \\ B_k - k_2 B_k & \text{if } 0\% < T_k^{\text{hypo}} \leq 5\% \\ B_k + a_k^b \, \mathrm{k}_l B_k, \forall a_k^b \epsilon \mathbf{A} & \text{if } T_k^{\text{hypo}} = 0\% \end{cases}$$

c. Apply B_{k+1}, observe the reward r_k^b and the next state $\mathbf{s_{k+1}^b}$.

d. Update the action value function (4).

e. $\mathbf{s_k^b} \leftarrow \mathbf{s_{k+1}^b}$.

$k_1 = 0.2$, $k_2 = 0.15$, and $k_1 = 0.1\left|\frac{G_k^{\text{mean}} - G_T}{10-4}\right|$ were chosen empirically using clinical guidelines and targeting an algorithm's convergence time in 7 iterations.

Overview of App Development

The mobile app was developed using JavaScript in React Native framework, allowing cross-platform compatibility with a single code. Firebase served as the back-end service, leveraging its encryption for enhanced security measures. The mobile app was developed for Android versions 6.0 and above (API level 23) as well as iOS versions 9.0 and higher. The mobile app was tested using Xray in Jira. Throughout the app development, GitHub was used for version control. TestFlight was used to distribute the mobile app to participants using iOS devices, while Firebase App Distribution was used to distribute the app to participants using Android devices. The reinforcement learning algorithms were implemented in MATLAB and ran on a Nitro 5 Core i7 computer. The weekly recommendations were sent to study participants using a cloud server. The app received Health Canada's Investigational Testing Authorization as a CLASS II medical device.

References

1 Sutton RS, Barto AG. Reinforcement learning: An introduction. *MIT Press* 2018.

2 Kok JR, Vlassis N. Sparse tabular multiagent Q-learning. *In Ann Mach Lear Conf of Belg and the Nether* 2004, pp. 65-71.

3 Vinyals O, Babuschkin I, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 2019; **575**: 350-354.

4 Zhang Z, Wang D. EAQR: A multiagent Q-learning algorithm for coordination of multiple agents. *Complexity* 2018; 2018: 1-4.

5 Kok JR, Vlassis N. Sparse tabular multiagent Q-learning. *In Annual Machine Learning Conference of Belgium and the Netherlands* 2004; 65-71. Enschede: Universiteit Twente Press.

6 Guestrin C, Venkataraman S, Koller D. Context-specific multiagent coordination and planning with factored MDPs. *In Proceedings of the Eighteenth National Conference on Artificial Intelligence* 2002; 253-259.

7 Kok JR, Vlassis N. Collaborative multiagent reinforcement learning by payoff propagation. *Journal of Machine Learning Research* 2006; **7**: 1789–1828.

8 Wiewiora E. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research* 2003;**19**: 205-208.